Week 4: beaks, death, and analyses

Silas Tittes

12 September, 2016

Outline

- Review of Grants' study (10 minutes)

Outline

- Review of Grants' study (10 minutes)
- Think, hypothesize, discuss (20 minutes)

Outline

- Review of Grants' study (10 minutes)
- Think, hypothesize, discuss (20 minutes)
- Review of functions, regression, and logistic regression (20 minutes)

Outline

- Review of Grants' study (10 minutes)
- Think, hypothesize, discuss (20 minutes)
- Review of functions, regression, and logistic regression (20 minutes)
- Lab (60 minutes)

Review of Grants' study

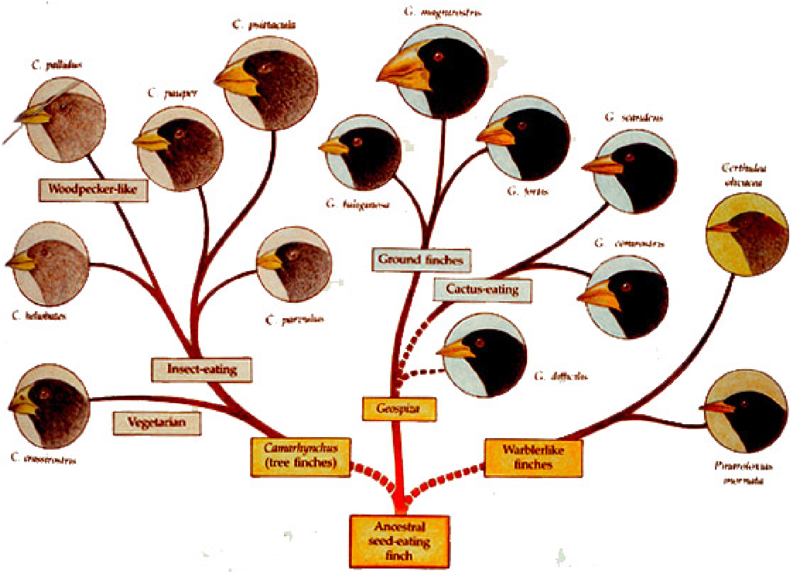


Figure 1: Darwin's Finches

Think, hypothesize, discuss

1. Propose a hypothesis about the effects of drought on the finches. Be sure to mention:
   -The phenotype of interest
   -Differential survival
2. Draw a graph of the predicted results of your hypothesis.
   - ☐ Axes are labeled
   - ☐ The predicted results stem directly from the hypothesis
   - ☐ The graph makes sense
   - ☐ It is ease to extract relevant information from the graph
3. Share with class

Think, hypothesize, discuss

4. Briefly describe how you would collect the data necessary to evaluate the hypothesis and test your prediction. Make your description a step-by-step set of procedures.
   ☐ The step-by-step procedures make sense
   ☐ The data that would come from the described procedures aligns with the predictions (in the graph above).

Think, hypothesize, discuss

Your boss (who just happens to be a world famous biologist) gives you some data and she asks you to assemble the data in a manner that can be used to assess whether selection happened during the drought on the island.

5. Construct a step-by-step algorithm of what you would do to assemble evidence that would allow you to make a claim about selection. The data are the sizes of finch beaks of individuals that survived and died. When you construct your algorithm, precede each statement with a hashtag(#). We'll get you started by providing two initial annotations and an annotation that will form the most relevant visualization.

Functions

Functions

Functions take input and return a new output Examples:

1. $$f(x) = x$$

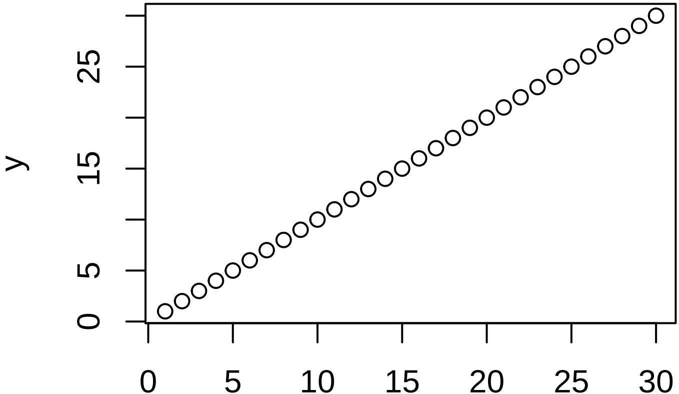
   - $$f(2) = 2$$

2. $$f(x) = x^2 + 1$$

   - $$f(2) = 2^2 + 1 = 5$$

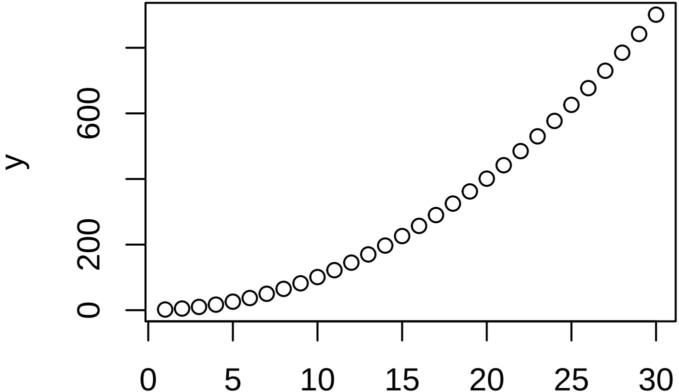
Function example 1

```r
fx <- function(x){
  x
}
x <- 1:30
y <- fx(x)
plot(x, y)
```



Function example 2

```
fx <- function(x){
  x^2 + 1
}
x <- 1:30
y <- fx(x)
plot(x, y)
```



Simple Linear Regression

Simple Linear Regression

- Uses a function that makes a straight line.

$$f(x, a, b) = a + bx$$

Simple Linear Regression

- Uses a function that makes a straight line.
- Adjust the parameters that control the height (intercept) and steepness (slope) of the straight line to find the best fit to a given dataset.

$$f(x, a, b) = a + bx$$

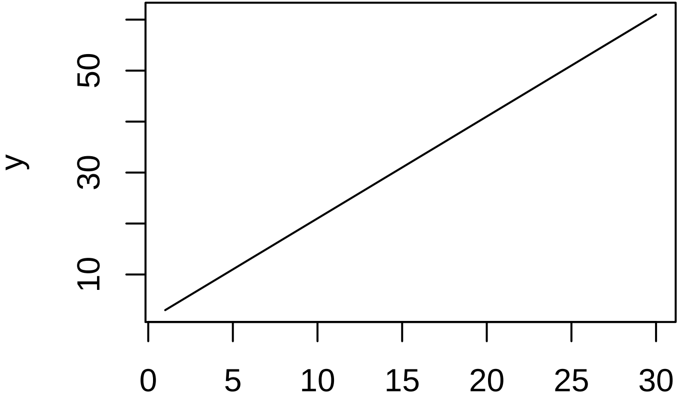
Simple Linear Regression

- Uses a function that makes a straight line.
- Adjust the parameters that control the height (intercept) and steepness (slope) of the straight line to find the best fit to a given dataset.
- Fancy math, beyond the scope of this course

$$f(x, a, b) = a + bx$$

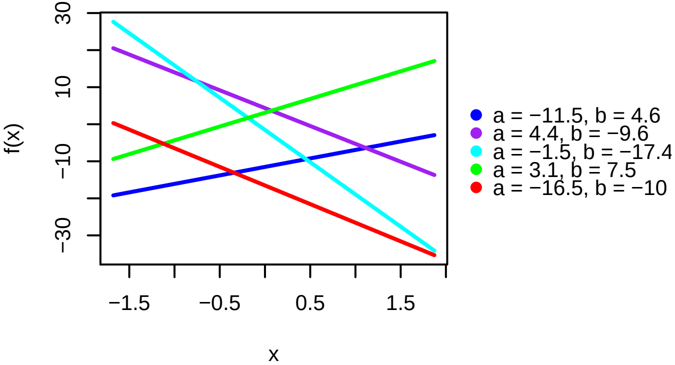
Simple Linear Regression

```
fx <- function(x, a, b){
  a + b*x
  }
y <- fx(x, a = 1, b = 2)
plot(x, y, type = "l")
```



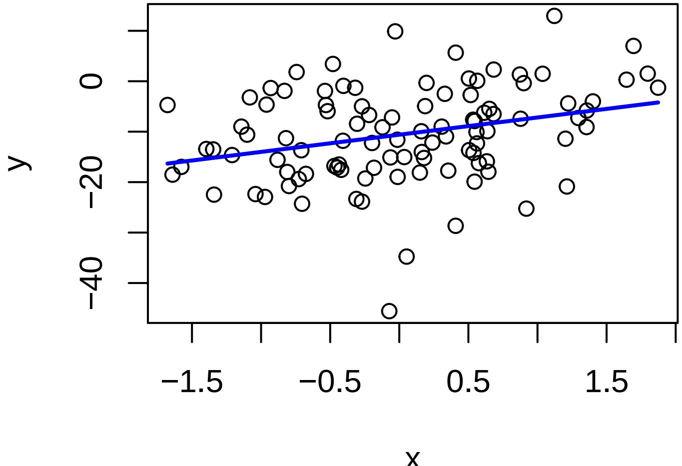
Examples of regression lines

$$f(x, a, b) = a + bx$$



With data

```
model <- lm(y ~ x)
predicted <- coef(model)[1] + coef(model)[2] * x
plot(x, y)
lines(x, predicted, lwd=2, col = "blue")
```



Reasons for doing regression

- To make predictions.

Reasons for doing regression

- To make predictions.
- To make quantitative and qualitative statements about relationships among variables.

Assumptions

- Among others, simple linear regression assumes the data is normally distributed.

Assumptions

- Among others, simple linear regression assumes the data is normally distributed.
- thus, y values should fall in range of $-\infty$ to ∞

Assumptions

- Among others, simple linear regression assumes the data is normally distributed.
- thus, y values should fall in range of $-\infty$ to ∞
- What if our y values are constrained? For example, what if our data is binary?

Logistic Regression

Logistic Regression

- Regression, but assuming our outcome variable (y) is binary (0 or 1, yes or no, black or white, dead or alive).

Logistic Regression

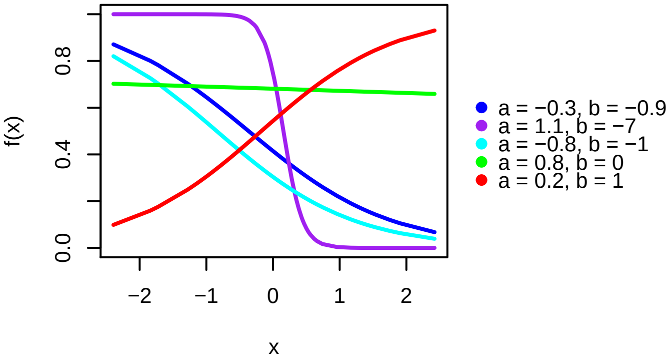
- Regression, but assuming our outcome variable (y) is binary (0 or 1, yes or no, black or white, dead or alive).
- Same as linear regression, but with a small twist that keeps predictions between 0 and 1.

Logistic Regression

- Regression, but assuming our outcome variable (y) is binary (0 or 1, yes or no, black or white, dead or alive).
- Same as linear regression, but with a small twist that keeps predictions between 0 and 1.
- Predictions are interpreted as probability of outcomes of a binary event occuring (probability of yes).

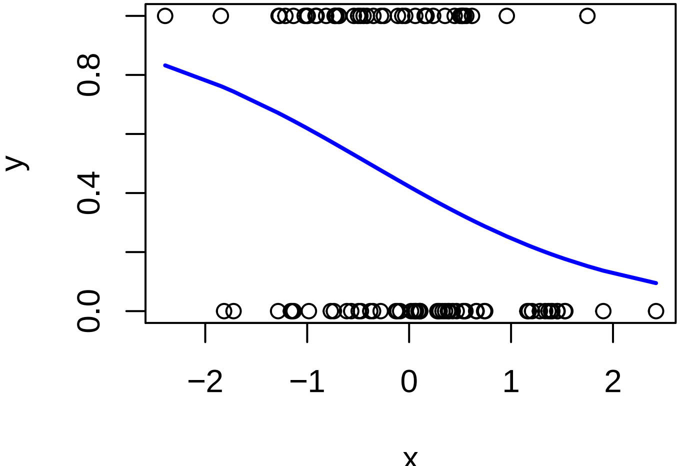
Examples of logistic regression curves

$$f(x, a, b) = \frac{e^{a+bx}}{e^{a+bx} + 1}$$



With data

```
model <- glm(y ~ x, family = "binomial")
predicted <- plogis(coef(model)[1] + coef(model)[2] * x)
plot(x, y)
lines(x, predicted, lwd=2, col="blue")
```



Note

- Logistic regression uses same principle's as linear regression, but is designed for binary data.

Note

- Logistic regression uses same principle's as linear regression, but is designed for binary data.
- Same mathematical machinery is used to find the best parameter values (a and b).

Note

- Logistic regression uses same principle's as linear regression, but is designed for binary data.
- Same mathematical machinery is used to find the best parameter values (a and b).
- Think it logistic regression as a transform of linear regression. Instead of intercept and slope of the line, we interpret parameters as position and steepness of transition between 0 and 1.

Today's lab

Today's lab

- .Rmd file on D2L

Today's lab

- .Rmd file on D2L
- Remember to put the .Rmd and .csv in the same place.

Today's lab

- .Rmd file on D2L
- Remember to put the .Rmd and .csv in the same place.
- Read data and use `str()` to understand data properties.